REPORT/WHITE PAPER

DATA SALARY PREDICTIVE MODEL

Joseph Choi | Jenny Overby | Daniel Meier | Serge Nane

DSC450 Applied Data Science
Spring 2024

TABLE OF CONTENTS

INTRODUCTION

In today's rapidly evolving job market, staying informed about salary trends is essential, especially for professionals in the data science and analytics field. As new and advanced technology emerges and companies adapt to these changes, the demand for skilled data professionals continues to grow, often leading to higher salaries. Understanding these trends is vital for new graduates to make informed decisions about job opportunities and salary negotiations. It also allows them to set realistic salary expectations when applying to roles. Therefore, developing a good grasp of the dynamics that influence salaries in these fields is necessary for informed career planning.

All members of our team are in the final semester of our data science bachelor's program. This project holds great significance to us as we'd like to gain a comprehensive understanding of salary trends in our field before entering the job market. With this in mind, we'd like to analyze salary trends using the most current data available for 2024. The dataset we are using is sourced from ai-jobs.net, providing detailed information on salary distribution globally. Please note that our project will focus primarily on salaries in the U.S., given its relevance to our future career prospects and job search efforts.

Here are the dataset's features and their descriptions:

- **work_year**: The year the data was recorded
- **job_title**: Specifies the job role
- **job_category**: Classification of the job role
- **salary_currency**: The currency in which the salary is paid
- **salary**: Annual gross salary in the local currency
- **salary_in_usd**: Annual gross salary converted to USD
- **employee_residence**: Residence of the employee
- **experience_level**: Professional experience level of the employee
- **employment_type**: Type of employment
- **work_setting**: Work environment
- **company_location**: The location of the company
- **company_size**: Size of the company

Link: [The Global AI, ML, Data Science Salary Index for 2024 | ai-jobs.net]

Through carefully executed data wrangling, analysis, science, and visualization, our team's main objective is to identify key factors influencing salary levels and create a predictive model based on the 2024 data.

BUSINESS PROBLEM

The core of our project revolves around two fundamental research questions:

1. What are the key factors dictating salary levels in data-related positions?
2. Can we construct a reliable predictive model that forecasts salaries using historical data?

Throughout the project, we centered our focus on these two main business problems, ensuring that all our data wrangling, science, and visualization tasks stayed aligned with these objectives to avoid going out of scope.

METHODS/ANALYSIS

To reach our objectives, our team divided the project into three segments. Here are the descriptions for each segment, detailing methods used and the insights extracted from our analysis:

1. DATA WRANGLING

We started our project by exploring and assessing the quality of our dataset. From the initial quality assessment, we identified several discrepancies that needed to be addressed before moving forward. The data cleaning and transformation tasks we performed included:

- **Corrected data types**: Converted "salary" from object to float by removing symbols
- **Handled missing values**: Replaced missing salary values with the mean
- **Addressed data entry errors**: Corrected zero values in "salary" based on "salary_in_usd"
- **Standardized categorical values**: Ensured consistency in capitalization and accuracy
- **Filtered non-US salary data**: Focused exclusively on US salary information
- **Reduced complexity by dropping unnecessary columns**:
    - Removed "company_location" and "employee_residence" for US-only focus
    - Eliminated "salary_currency" and "salary_in_usd" (redundant information

2. DATA SCIENCE

Once we confirmed that the dataset was ready for further analysis and model training, we transitioned to the data science portion of the project. This segment had two phases: EDA and Model training.

EXPLORATORY DATA ANALYSIS (EDA)

For EDA, our objective was to understand the characteristics of each input feature and our target variable. The focus was to extract insights to help guide and develop an approach to our feature selection and feature engineering process.

Here is a list of tasks performed during the EDA phase:

- Built a histogram and box plot for our target variable to analyze its distribution and identify outliers
- Built bar charts comparing the average salary with each input categorical feature to understand their distributions and identify any dominant categories that might impact the learning process
- Built box plots comparing the salary with each input categorical feature to identify outliers that could affect model training
- Built a correlation matrix to comprehend the correlation coefficients between all variables. The output, however, was not helpful as the matrix failed to capture the relationship between our input features (categorical and encoded) and target variable (numerical).

Insights extracted and decisions made from our EDA relating to our feature selection, feature engineering, and model training are as below:

- The histogram of the salary column displays a bell-shaped curve (close to), indicating a normal distribution (Fig. 1). This suggests that linear regression models are advantageous due to their compatibility with normally distributed target variables. Therefore, for our model selection, we'd prioritize linear regression models like ridge and lasso for our training process. In addition,

we are adding non-linear regression models (Random Forest and Gradient Boosting Machines) to assess their performance with our dataset.

Fig. 1: Salary Histogram

- Box plots displayed outliers past the $300,000 mark, indicating salaries significantly higher than the majority (Fig. 2). After a thorough investigation, we concluded that these outliers were legitimate and decided to keep them in our model training process. However, in one of our model training iterations, we plan to cap these outliers at $310,000 (value just before the outlier threshold) to assess their impact on model performance.

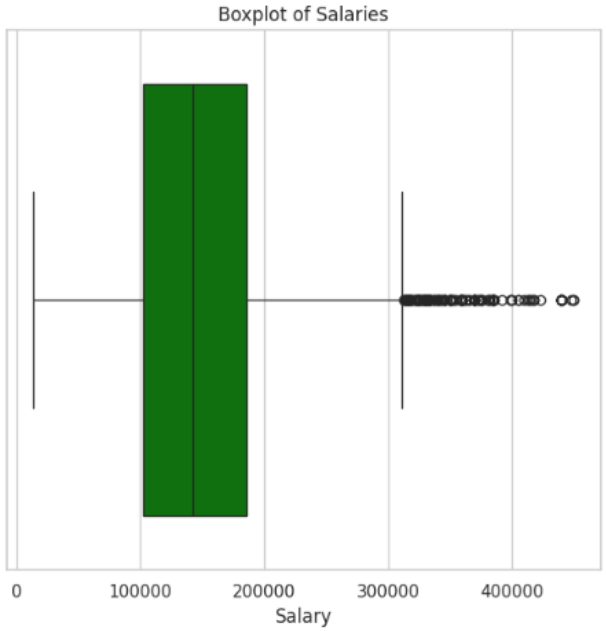

Fig. 2: Salary Boxplot

- Bar charts showed no signs of dominant categories in the dataset, which is essential to avoid bias in the model training process. Each category per feature displayed a balanced progression of increase and decrease, indicating a healthy distribution.
- We decided to remove the job title column as it was redundant with the job category column. The job title column contained over 150 unique values that would not have positively contributed to model training. Therefore, the column was removed to simplify the dataset.
- For encoding, we are using ordinal encoding for work year, experience level, employment type, work setting, and company size to assign numbers to each feature based on order. We are using one-hot encoding for the job category to treat each category independently without assuming any ordinal relationship.
- As our correlation matrix failed to capture the correlation coefficients properly, we have decided to pull feature importance information from our linear and non-linear regression models using scikit-learn. The coefficient outputs will serve as our main indicator in identifying variables with a high impact on our target variable, salary.

MODEL TRAINING

Once the dataset was prepared for model training, we experimented with four models (Ridge, Lasso, Random Forest, Gradient Boosting Machines), running three iterations each. We documented and evaluated performance metrics for each iteration, selecting the best-performing model as our final choice. We made intentional adjustments throughout each iteration based on a predefined plan:

1. In the first iteration, all selected features for the model training process were included.
2. In the second iteration, outliers in the salary column were addressed by capping extreme values at $310,000, replacing values beyond this threshold.
3. In the third iteration, we included only a select few features in the training process. Specifically, we included the top 3 features with the highest predictive power, mainly determined by the feature importance bar chart that was developed for each model type (Fig. 3).

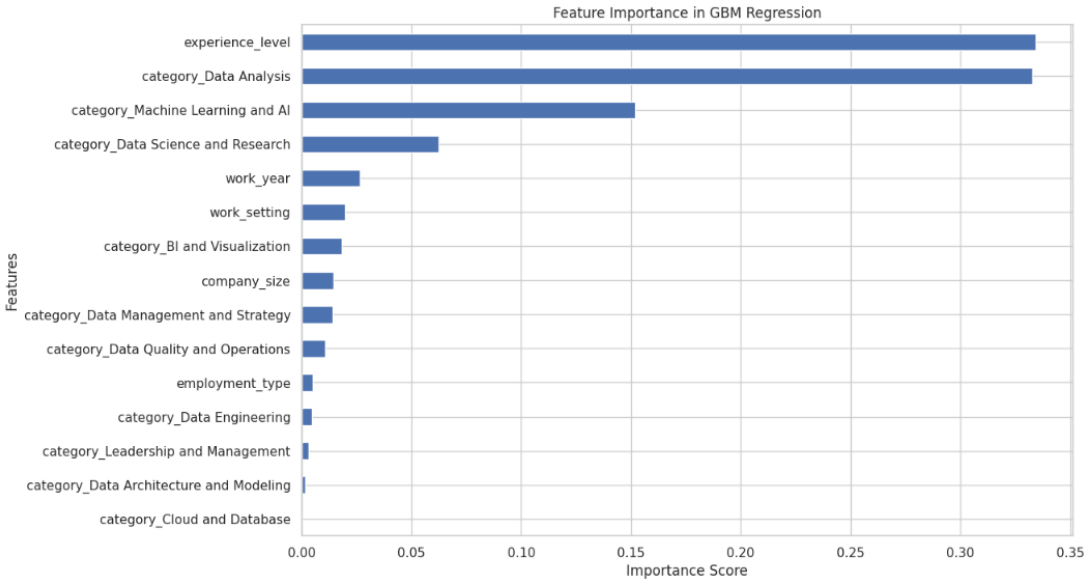

Fig. 3: GBM Feature Importance - Bar Graph

After 12 iterations, the best-performing model was chosen as our final selection. Only to this final model, we adjusted the hyperparameters using grid search to optimize the performance of the model.

3. DATA VISUALIZATION

While the data science phase was happening, our team was simultaneously creating data visualizations to support our analysis and model training process. We wanted a deeper understanding of the dataset and leverage these visuals to make informed decisions about addressing our research questions.

Out of all the data visualizations we generated, one particular visual that stood out was the distribution plot that displays the salary distribution across different job categories, segmented by work years. Each panel, differentiated by year, highlights trends over time to quickly scan how the job market is changing annually. Based on the plot, we noticed a significantly low job volume in 2020 compared to other years. Then as the years progressed, the job volume gradually increased. By 2023, the job market experienced a significant increase, led by a surge in job opportunities (Fig. 4).

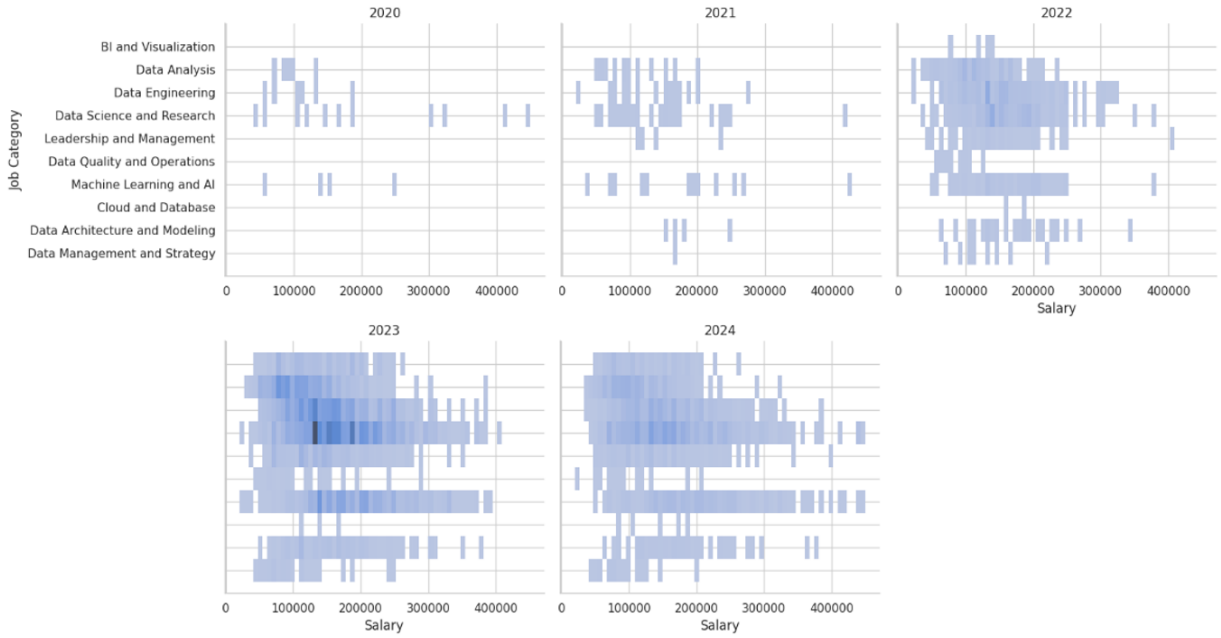

Fig. 4: Salary by Year and Category - Distribution Plot

A similar trend was displayed when observing other features (experience level and work setting), in place of the job category feature (Fig. 5).

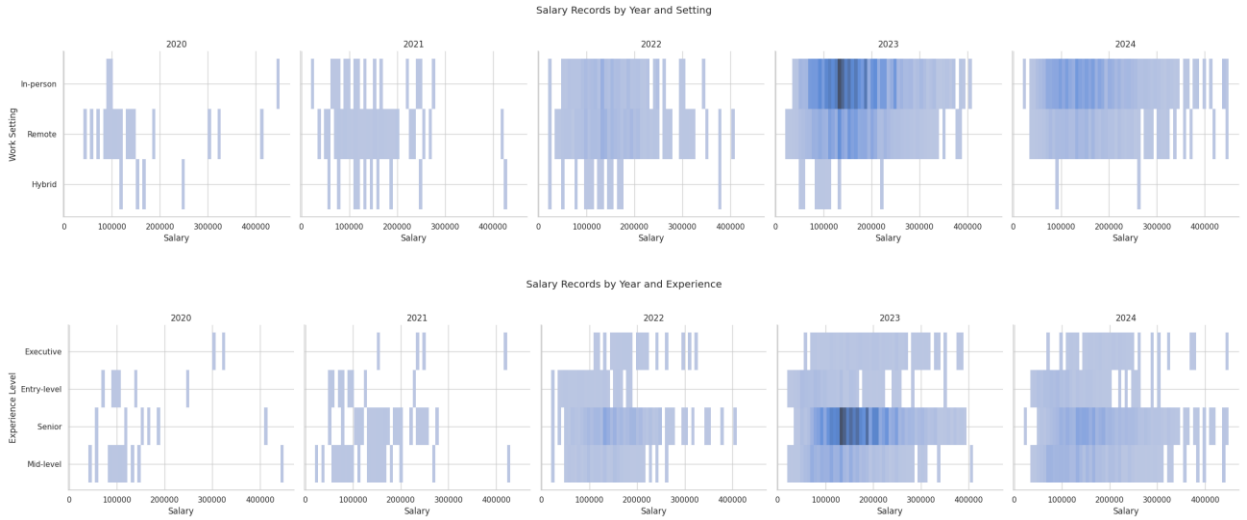

Fig. 5: Salary by Year and Work Setting/Experience Level - Distribution Plots

RESULTS

Referencing our GBM Feature Importance Bar Graph visual, the key variables that impact salary are experience level, job categories (specifically data analysis, machine learning and AI, data science and research), and work year.

We evaluated many different regression models and iterations to predict salaries, primarily using the R2 score as our performance metric. Here are the results from our model training and evaluation phase:

- **Ridge**: This model showed a limited ability to predict salaries with an R2 score of 0.2498
- **Lasso**: This model performed slightly better than Ridge with an R2 score of 0.2622
- **Random Forest**: This model performed slightly better than Ridge but fell short compared to Lasso with an R2 score of 0.2588
- **Gradient Boosting Machines**: This model had the best results with an R2 score of .2680
- **Hyperparameter Tuning:** Post hyper-tuning, the GBM model showed little to no improvement with an R2 value of 0.2672

Based on our evaluation and tuning results, we found that the GBM model showed the best performance out of all models. From our iterations, we noticed a decrease in performance for all models when we simplified them to include only the top three features. This indicates the need for a broader range of features to capture the complexities of our dataset effectively. Expanding on this, our final model did not perform as well as we expected. We concluded that the reason for the low R2 score is due to the dataset lacking essential information for accurate salary prediction. We believe that adding additional features, such as economic indicators like inflation and unemployment rates, could improve the model's performance. Currently, the model is unable to capture these economic trends using the dataset we have.

In addition, considering the dataset that we are plugging into our models, most input features are encoded categorical values, while the target variable is numeric. Putting this into context, we recognize the need to improve our feature engineering process to optimize how our models can learn from the dataset. We also acknowledge that the model selection could be better to be more suited to the current data structure.

Detailed documentation of the model training for each iteration can be found in the appendix.

NEXT STEPS

Based on our outcomes and insights derived from our initial model training and evaluation, our next steps for the project would be to focus on improving the predictive accuracy of our salary prediction models through these steps:

1. Our analysis pointed out the potential for improvement through the addition of more features related to economic indicators and trends. We plan to find data sources that provide these metrics and integrate them into our dataset.
2. We plan to explore more advanced feature engineering techniques. In this step, our main focus would be to find better ways to encode our data that preserve crucial information from our categorical features and establish a strong connection with our numerical target variable, salary.
3. We also plan to explore other models that may better suit our feature set. We would like to test the abilities of these alternative models to see how well they perform against our initial models.
4. We intend to conduct further research and perform a more thorough hyperparameter tuning process to optimize our models for improved performance.
5. We would like to apply cross-validation techniques to better assess our model's generalizability to ensure it performs well across various subsets of data. This approach will help us understand how our models can adapt to changes in economic conditions.

CONCLUSION

Bringing everything together, our team applied various data science methodologies to analyze and predict salary trends in today's evolving data job market. Using the data extracted from ai-jobs.net, our analysis identified key factors affecting salaries and showed the Gradient Boosting Machines (GBM) model as our top performer. However, our models fell short and did not perform as well as we expected. To address this, our plan moving forward is to enhance our dataset, explore advanced data science techniques for feature engineering, model selection, and hyperparameter tuning, and implement cross-validation to strengthen our model reliability. Ultimately, our goal is to improve our predictive capabilities to empower new graduates and professionals to make informed career and salary decisions.

REFERENCES

Camm, J. D., Bowers, M. R., & Davenport, T. H. (2020, June 16). The Recession's Impact on Analytics and Data Science. MIT Sloan Management Review. https://sloanreview.mit.edu/article/the-recessions-impact-on-analytics-and-data-science/

Dialani, P. (2020, July 14). Will Recession Impact Data Science and Analytics? Analytics Insight. https://www.analyticsinsight.net/will-recession-impact-data-science-and-analytics/

Kablaoui, R., & Salman, A. (2022, November 23). Machine Learning Models for Salary Prediction Dataset using Python. In International Conference on Electrical and Computing Technologies and Applications (ICECTA). Semanticscholar. https://www.semanticscholar.org/paper/Machine-Learning-Models-for-Salary-Prediction-using-Kablaoui-Salman/5a213573154231fe97113f1c41bc3651a3eea409

Lai, J. (2020, October 12). Predict Data Science Salaries with Data Science: Dissecting the US Data Science job market and salaries in 2020 with Exploratory Analysis and Machine Learning algorithms in Python. Towards Data Science. https://towardsdatascience.com/the-us-data-science-job-market-in-2020-463520a9d5a

Ma, B. (2020, June 19). Job Salary Prediction with NLP, Machine Learning and Deep Learning. Towards Data Science. https://towardsdatascience.com/job-salary-prediction-with-nlp-machine-learning-and-deep-learning-b87a96336b08

Saish, A. (2023, February 23). A data-driven approach to salary prediction. Medium. https://medium.com/@mummineniamar/a-data-driven-approach-to-salary-prediction-158f6cb8f121

Sethuramaswamy, S. (2022, December 15). Key Data And Analytics Trends To Watch In 2023. Forbes Technology Council. https://www.forbes.com/sites/forbestechcouncil/2022/12/15/key-data-and-analytics-trends-to-watch-in-2023/?sh=1b65271629b0

Tee, Z., & Raheem, M. (2022, July). Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits: A Literature Review. ResearchGate. https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review

Wang, G. (2022). Employee Salaries Analysis and Prediction with Machine Learning. In 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) (pp. 373-378). Guangzhou, China. https://doi.org/10.1109/MLISE57402.2022.00081

Yosifova, A. (2024, April 11). The Data Scientist Job Market in 2024 [Research on 1,000 Job Postings]. 365 Data Science. https://365datascience.com/career-advice/data-scientist-job-market/

APPENDIX

Detailed model training iterations and corresponding R2 values:

RIDGE REGRESSION RESULTS:

1. **Full Model with All Features**
    - **RMSE**: High at 53059.834240 | **R^2 Score**: Low at 0.2498462266
2. **Outlier Adjustment**
    - **RMSE**: Unchanged (same as the 1st) | **R^2 Score**: Unchanged (same as the 1st)
3. **Top 3 Features Model**
    - **Selected Features**: 'category_machine learning and AI', 'experience_level', 'category_data_science_and_research'
    - **RMSE**: Increased to 54160.1960 | **R^2 Score**: Decreased to 0.2184100

Summary: Ridge Regression showed limited predictive ability, with a low R^2 score indicating that the model captured a small proportion of the variance in salaries. Simplifying the model to the top 3 features worsened the performance.

LASSO REGRESSION RESULTS:

1. **Full Model with All Features**
    - **RMSE**: 50540.5603 | **R^2 Score**: 0.2622
2. **Outlier Adjustment**
    - **RMSE**: Unchanged (same as the 1st) | **R^2 Score**: Unchanged (same as the 1st)
3. **Top 3 Features Model**
    - **Selected Features**: 'experience_level', 'employment_type', 'category_machine learning_and_AI'
    - **RMSE**: Increased to 51675.3532 | **R^2 Score**: Decreased to 0.2287

Summary: LASSO Regression had a slightly better R^2 score compared to Ridge Regression, indicating a marginally better fit. However, it still demonstrated a weak predictive performance. The performance decreased when the model was simplified, suggesting that other features are necessary for a more accurate prediction.

RANDOM FOREST REGRESSION RESULTS:

1. **Full Model with All Features**
    - **RMSE**: 50654.6217 | **R^2 Score**: 0.2588

2. **Outlier Adjustment**

    - **RMSE**: Unchanged (same as the 1st) | **R^2 Score**: Unchanged (same as the 1st)

3. **Top 3 Features Model**

    - **Selected Features**: 'experience_level', 'category_data_analysis', 'work_year'
    - **RMSE**: Increased to 51472.0897 | **R^2 Score**: Decreased to 0.2347

Summary: Random Forest Regression showed an improvement in R^2 score over the Ridge models but still was not highly predictive. The performance decreased when the model was simplified, suggesting that other features are necessary for a more accurate prediction.

GRADIENT BOOSTING MACHINES REGRESSION RESULTS:

1. **Full Model with All Features**

    - **RMSE**: 50341.3481 | **R^2 Score**: 0.2680

2. **Outlier Adjustment**

    - **RMSE**: Unchanged (sale as the 1st) | **R^2 Score**: Unchanged (sale as the 1st)

3. **Top 3 Features Model**

    - **Selected Features**: 'experience_level', 'category_data analysis', 'category_machine_learning_and_AI'
    - **RMSE**: Increased to 51468.5871 | **R^2 Score**: Decreased to 0.2348

Summary: GBM Regression demonstrated the best initial performance among the models tested, with the lowest RMSE and the highest R^2 score. However, this performance dropped when the model was reduced to the top 3 features, suggesting the need for a broader feature set.

Conclusion: The models tested had varying degrees of success in predicting salaries, with GBM showing the best initial results. Simplification to the top features generally reduced model performance, indicating that salary predictions benefit from a broader range of inputs. The unchanged metrics after outlier adjustment across all models suggest a limited impact of outliers on predictions, possibly due to a balanced distribution of such values across the dataset or their relatively small number. To enhance predictive power, we will model tuning.

HYPERPARAMETER TUNING:

Hyper-tuning the GBM Regression Model:

- Fitting 3 folds for each of 972 candidates, totaling 2916 fits
- **Optimized RMSE**: 50368.8279
- **Optimized R^2 Score**: 0.2672
- **Best hyperparameters**: {'learning_rate': 0.1, 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100, 'subsample': 0.8}

Hyper-tuning the Lasso Regression Model:

- **Optimized RMSE**: 50540.5602
- **Optimized R^2 Score**: 0.2622
- **Best alpha**: 10.0